Summary

Analysis for X Education and we have found new ways to get more industry professionals to join their courses. The data provided by the company gave us a lot of information about how the potential customers approaching website, the time spent, medium through which they find the site and the conversion rate. Also, we did analysis on the data and found new outcomes

The following are the steps used:

1. **Cleaning data:**
   The data was cleaned as there were values as “select” in many rows so that value we considered as null as there was no any information provided on it. The null values were changed to ‘nan’ so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to ‘India’, ‘Foreign Country’ and ‘Missing Value’.

2. **EDA:**
   We perform EDA to check the condition of our data. The numeric values seems good and no outliers were found. A lot of elements in the categorical variables were irrelevant.

3. **Creating Dummy Variables:**
   The dummy variables were created and later on the dummies with “Missing Values” elements were removed. For numeric values we used the MinMaxScaler.

4. **Test-Train split:**
   The split was done at 70% and 30% for train and test data respectively.

5. **Model Building:**
   Firstly, we imported Logistic Regression after that we perform RFE to attain the top 15 relevant variables. Later the rest of thevariables were removed manually depending on the VIF values and p-value (Thevariables with VIF < 5 and p-value < 0.05 were kept).

6. **Prediction on train data set**
   Prediction was done on the train data frame and with an optimum cut off as 0.5

7. **Model Evaluation:**
   We have created confusion matrix. Later the optimum cut off value (using ROC curve)was used to find the accuracy, sensitivity and specificity which came to be around 81% each.

8. **Prediction on test data set**
   Prediction was done on the test data frame and with an optimum cut off as 0.5 withaccuracy, sensitivity and specificity of 80%.

9. **Precision – Recall:**
   This method was also used to recheck and a cut off of 0.81 was found with Precision around 74% and recall around 75%.

Conclusion

We have found that the variables that mattered the most in the potential buyers are (In descending order) :

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was
   a. Google
   b. Direct traffic
   c. Organic search
   d. Welingak website
4. When the last activity was
   a. SMS
   b. Olark chat conversation
5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional. Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.